What is Multivariate Regression Analysis and how can it be used in Stata for Data Analysis?

Authored by

stats writer

June 29, 2024

Multivariate Regression Analysis is a statistical technique used to examine the relationship between multiple independent variables and a single dependent variable. It is a powerful tool for analyzing complex data sets and identifying significant factors that influence the outcome variable. In Stata, this analysis can be performed by using the "regress" command, which allows for the inclusion of multiple independent variables and provides various diagnostic tests to evaluate the model's validity. Multivariate Regression Analysis in Stata can be used for data analysis in various fields, such as economics, social sciences, and business, to understand the impact of multiple factors on a particular outcome and make informed decisions based on the results.

Multivariate Regression Analysis | Stata Data Analysis Examples

Version info: Code for this page was tested in Stata 12.

As the name implies, multivariate regression is a technique that estimates a single regression model with more than one outcome variable. When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.

Please Note: The purpose of this page is to show how to use various data analysis commands. It does not cover all aspects of the research process which researchers are expected to do. In particular, it does not cover data cleaning and checking, verification of assumptions, model

diagnostics and potential follow-up analyses.

Examples of multivariate regression

Example 1. A researcher has collected data on three psychological variables,
four academic variables (standardized test scores), and the type of educational
program the student is in for 600 high school students. She is interested in how
the set of psychological variables is related to the academic variables and the
type of program the student is in.

Example 2. A doctor has collected data on cholesterol, blood pressure, and
weight. She also collected data on the eating habits of the subjects
(e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed
per week). She wants to investigate the relationship between the three
measures of health and eating habits.

Example 3. A researcher is interested in determining

what factors influence
the health African Violet plants. She collects data on the average leaf
diameter, the mass of the root ball, and the average diameter of the blooms, as
well as how long the plant has been in its current container. For predictor variables,
she measures several elements in the soil, as well as the amount of light
and water each plant receives.

Description of the data

Let's pursue Example 1 from above. We have a hypothetical dataset with 600
observations on seven variables. he psychological variables are locus of control
(locus_of_control), self-concept (self_concept), and motivation (motivation). The academic variables are standardized tests scores in
reading (read), writing (write), and science (science), as well as a categorical
variable (prog) giving the type of program the student is in (general,

academic, or vocational).

Let's look at the data (note that there are no missing values in this data set).

use https://stats.idre.ucla.edu/stat/stata/dae/mvreg, clearsummarize locus_of_control self_concept motivation read write science

Variable | Obs Mean Std. Dev. Min Max
-------------+--
locus_of_c~l | 600 .0965333 .6702799 -1.995957 2.205511
self_concept | 600 .0049167 .7055125 -2.53275 2.093563
motivation | 600 .0038979 .8224 -2.746669 2.583752
read | 600 51.90183 10.10298 24.62001 80.58649
write | 600 52.38483 9.726455 20.06888 83.93482
-------------+--
science | 600 51.76333 9.706179 21.98953 80.36942

tabulate prog

program |
type | Freq. Percent Cum.
------------+-----------------------------------
general | 138 23.00 23.00

academic | 271 45.17 68.17
vocational | 191 31.83 100.00
----------+----------
Total | 600 100.00

correlate locus_of_control self_concept motivation
(obs=600)

| locus_~l self_c~t motiva~n
----------+----------
locus_of_c~l | 1.0000
self_concept | 0.1712 1.0000
motivation | 0.2451 0.2886 1.0000

correlate read write science
(obs=600)

| read write science
----------+----------
read | 1.0000
write | 0.6286 1.0000
science | 0.6907 0.5691 1.0000

Analysis methods you might consider

Below is a list of some analysis methods you may have

encountered.
Some of the methods listed are quite reasonable while others have either
fallen out of favor or have limitations.

Multivariate regression

To conduct a multivariate regression in Stata, we need to use two commands,
manova and mvreg. The manova command will indicate if
all of the equations, taken together, are statistically significant. The
F-ratios and p-values for four
multivariate criterion are given, including Wilks' lambda, Lawley-Hotelling
trace, Pillai's trace, and Roy's largest root. Next, we use the mvreg
command to obtain the coefficients, standard errors, etc., for each of the predictors in
each part of the
model. We will also show the use of the test command after the
mvreg command. The use of the test command is one of the

compelling reasons for conducting a multivariate regression analysis.

Below we run the manova command. Note the use of c. in front of the
names of the continuous predictor variables -- this is part of the factor variable
syntax introduced in Stata 11. It is necessary to use the c. to identify
the continuous variables, because, by default, the manova command assumes all
predictor variables are categorical.

manova locus_of_control self_concept motivation = c.read c.write c.science prog

Number of obs = 600

W = Wilks' lambda L = Lawley-Hotelling trace
P = Pillai's trace R = Roy's largest root

Source | Statistic df F(df1, df2) = F Prob>F
----------+----------
Model | W 0.6992 5 15.0 1634.7 15.08 0.0000 a
| P 0.3196 15.0 1782.0 14.17 0.0000 a

| L 0.4035 15.0 1772.0 15.89 0.0000 a

| R 0.3271 5.0 594.0 38.86 0.0000 u

|----------

Residual | 594

----------+----------

read | W 0.9764 1 3.0 592.0 4.76 0.0027 e

| P 0.0236 3.0 592.0 4.76 0.0027 e

| L 0.0241 3.0 592.0 4.76 0.0027 e

| R 0.0241 3.0 592.0 4.76 0.0027 e

|----------

write | W 0.9474 1 3.0 592.0 10.96 0.0000 e

| P 0.0526 3.0 592.0 10.96 0.0000 e

| L 0.0555 3.0 592.0 10.96 0.0000 e

| R 0.0555 3.0 592.0 10.96 0.0000 e

|----------

science | W 0.9834 1 3.0 592.0 3.33 0.0193 e

| P 0.0166 3.0 592.0 3.33 0.0193 e

| L 0.0169 3.0 592.0 3.33 0.0193 e

| R 0.0169 3.0 592.0 3.33 0.0193 e

|----------

prog | W 0.8914 2 6.0 1184.0 11.67 0.0000 e

| P 0.1086 6.0 1186.0 11.35 0.0000 a

| L 0.1217 6.0 1182.0 11.99 0.0000 a

| R 0.1209 3.0 593.0 23.89 0.0000 u

|----------

Residual | 594

----------+----------

Total | 599

----

e = exact, a = approximate, u = upper bound on F

We can use mvreg to obtain estimates of the coefficients in our model.
Normally mvreg requires the user to specify both outcome and predictor
variables, however, because we have just run the manova command, we can use the mvreg command, without
additional input, to run a multivariate regression corresponding to the model just
estimated by maova (note that this feature was introduced in Stata 11, if
you are using an earlier version of Stata, you'll need to use the full syntax for mvreg).

mvreg

Equation Obs Parms RMSE "R-sq" F P

---

locus_of_c~l 600 6 .6069966 0.1868 27.28199 0.0000
self_concept 600 6 .6890684 0.0540 6.786094 0.0000
motivation 600 6 .761402 0.1500 20.96388 0.0000

| Coef. Std. Err. t P>|t|
---+---
locus_of_c~l |
read | .0125046 .0037178 3.36 0.001 .005203 .0198062
write | .012145 .0033914 3.58 0.000 .0054845 .0188056
science | .0057615 .0036412 1.58 0.114 -.0013896 .0129126
|
prog |
2 | .1277951 .063955 2.00 0.046 .0021896 .2534005
3 | .2516705 .0684699 3.68 0.000 .117198 .386143
|
_cons | -1.624765 .1570053 -10.35 0.000 -1.933118 -1.316412
---+---
self_concept |
read | .0013076 .0042205 0.31 0.757 -.0069812 .0095965
write | -.0042934 .0038499 -1.12 0.265 -.0118545 .0032676

science | .0053059 .0041335 1.28 0.200 -.0028121 .013424

|

prog |

2 | .2764834 .0726023 3.81 0.000 .1338949 .4190719

3 | .4233592 .0777277 5.45 0.000 .2707047 .5760137

|

_cons | -.3723412 .1782339 -2.09 0.037 -.7223865 -.0222958

-------------+--

motivation |

read | .0096735 .0046635 2.07 0.038 .0005146 .0188325

write | .0175354 .004254 4.12 0.000 .0091807 .0258902

science | -.0090015 .0045674 -1.97 0.049 -.0179716 -.0000313

|

prog |

2 | .3603294 .0802236 4.49 0.000 .2027729 .5178859

3 | .619696 .085887 7.22 0.000 .4510169 .7883751

|

_cons | -1.310842 .1969437 -6.66 0.000 -1.697633 -.9240513

If you ran a separate OLS regression for each outcome variable, you would get exactly the same coefficients, standard errors, t- and p-values, and confidence intervals as shown above. So why conduct a multivariate regression? As we mentioned earlier, one of the advantages of using mvreg is that you can conduct tests of the coefficients across the different outcome variables. (Please note that many of these tests can be preformed after the manova command, although the process can be more difficult because a series of contrasts needs to be created.) In the examples below, we test four different hypotheses.

For the first test, the null hypothesis is that the coefficients for the variable read are equal to 0 in all three equations. (Note that this duplicates the test for the variable read in the manova output above.)

test read

(1) read = 0

(2) read = 0

(3) read = 0

F(3, 594) = 4.78

Prob > F = 0.0027

The results of this test reject the null hypothesis that the coefficients for
read across the three equations are simultaneously equal to 0, in other
words, the coefficients for read, taken for all three outcomes together,
are statistically significant.

Second, we can test the null hypothesis that the coefficients for prog=2
(identified as 2.prog) and prog=3 (identified as 3.prog) are simultaneously equal to 0 in the
equation for
locus_of_control. When used to test the coefficients for dummy variables
that form a single categorical predictor, this type of test is sometimes called an overall test

for the effect of the categorical predictor (i.e. prog). Note that the variable name in brackets (i.e. locus_of_control) indicates which equation the coefficient being tested
belongs to, with the equation identified by the name of the outcome variable.

test 2.prog 3.prog

(1) 2.prog = 0
(2) 3.prog = 0

F(2, 594) = 6.83
Prob > F = 0.0012

The results of the above test indicate that the two coefficients together are
significantly different from 0, in other words, the overall effect of prog
on locus_of_control
is statistically significant.

The next example tests the null hypothesis that the coefficient for the variable
write in the equation with

locus_of_control as the outcome is equal to the coefficient for write
in the equation with self_concept as the outcome. The null hypothesis
printed by the test command is that the difference in the coefficients is 0,
which is another way of saying two coefficients are equal. Another way of
stating this null hypothesis is that,
that the effect of write on locus_of_control is equal to the
effect of write on self_concept.

```
test write = write

( 1) write - write = 0

F( 1, 594) = 11.89
Prob > F = 0.0006
```

The results of this test indicate that the difference between the
coefficients for write with locus_of_control and
self_concept as the outcome is significantly different from 0, in other

words, the coefficients are significantly different.

For the final example, we test the null hypothesis that the
coefficient of science in the equation for
locus_of_control is equal to the coefficient for science in the
equation for self_concept, and that the coefficient for the variable
write in the equation with the outcome variable
locus_of_control equals the coefficient for write in the
equation with the outcome variable self_concept. We tested the
difference in the coefficients for write in the last example, so we can use
the accum option to add the test of the difference in coefficients
for science, allowing us to test both sets of coefficients at the
same time.

```
test science = science, accum

( 1) write - write = 0
```

(2) science - science = 0

F(2, 594) = 6.39
Prob > F = 0.0018

The results of the above test indicate that taken together the differences in the two sets of coefficients is statistically significant.

Things to consider

See Also

Stata Online Manual

References